Probablità, Statistica e Processi Stocastici

Franco Flandoli, Università di Pisa

Corso per la Scuola di Dottorato in Ingegneria

Matrice di covarianza

Dato un vettore aleatorio (anche non gaussiano) $\mathbf{X} = (X_1, ..., X_n)$, chiamiamo sua media il vettore

$$\boldsymbol{\mu} = (E[X_1], ..., E[X_n])$$

e matrice di covarianza la matrice $n \times n$ di componenti

$$Q_{ij} = Cov(X_i, X_j), \qquad i, j = 1, ..., n$$

Ricordiamo che

$$\begin{aligned} Cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

e che questa operazione è lineare in entrambi i sui argomenti.

Theorem

La matrice Q è simmetrica, definita non negativa.

Decomposizione spettrale

- Essendo simmetrica, Q è diagonalizzabile: esiste una base ortonormale $\mathbf{e}_1, ..., \mathbf{e}_n$ di $\mathbb{R}^n$ fatta di autovettori di Q,

$$Q\mathbf{e}_i = \lambda_i \mathbf{e}_i$$

e, posto

$$Q_e = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}, \qquad U = \begin{pmatrix} \mathbf{e}_1 & \dots & \mathbf{e}_n \\ & & \end{pmatrix}$$

vale

$$Q = U Q_e U^T.$$

- Essendo Q positiva, gli autovalori λ_i risultano positivi.

Definizione di radice quadrata

Ricordando $Q = UQ_eU^T$, basta ora porre

$$\sqrt{Q} = U\sqrt{Q_e}U^T.$$

Si verifica subito che $\sqrt{Q}\sqrt{Q} = Q$ e che $\sqrt{Q}$ è simmetrica e definita positiva.
Ecco i comandi di R:

```
e = eigen(Q)
U = e$vectors
B = U %*% diag(sqrt(e$values)) %*% t(U)
```

Esercizio riassuntivo

Svolgiamo insieme un esercizio riassuntivo.

1. Creiamo un vettore di punti gaussiani generati con R, che corrispondano ad una gaussiana dilatata in una direzione e ruotata di $\pi/10$
2. Di tali punti calcoliamo la matrice di covarianza empirica, che chiameremo Q
3. Di Q calcoliamo la radice quadrata $\sqrt{Q}$
4. Usando $\sqrt{Q}$, simuleremo punti gaussiani aventi covarianza Q
5. osserveremo ad occhio la somiglianza coi punti creati all'inizio (volendo ci sarebbero dei test statistici)

L'esercizio si trova svolto in una scheda a parte.

Verso PCA

- Supponiamo di avere dei punti nello spazio 3D.
- Ad esempio, supponiamo che siano stati generati da una gaussiana tridimensionale.
- Se vogliamo vederli nel modo più "aperto" ("sparpagliato") possibile, meno "sovrapposto", come dobbiamo ruotare il disegno?
- Immaginate che i punti siano una nuvola ellissoidale; l'ellissoide avrà un asse più lungo, un secondo asse più lungo ed un terzo asse più corto. Se ci mettiamo nel piano dei due assi più lunghi abbiamo la visuale più sparpagliata possibile.
- Come si trovano gli assi dell'ellisoide?

Le superfici di livello di una gaussiana qualsiasi

Theorem

Supponiamo $\det Q \neq 0$. *Allora le superfici di livello della densità*

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det Q}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T Q^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

ovvero gli insieli della forma

$$\left\{\mathbf{x} \in \mathbb{R}^n : (\mathbf{x}-\boldsymbol{\mu})^T Q^{-1}(\mathbf{x}-\boldsymbol{\mu}) = r_a^2\right\}$$

sono ellissoidi aventi come assi gli autovettori $\mathbf{e}_1, ..., \mathbf{e}_n$ *di* Q, *e lunghezze degli assi i numeri* $\sqrt{\lambda_1}, ..., \sqrt{\lambda_n}$, *dove* $\lambda_1, ..., \lambda_n$ *sono gli autovalori corrispondenti a* $\mathbf{e}_1, ..., \mathbf{e}_n$. *Useremo la convenzione che sia* $\lambda_1 \geq ... \geq \lambda_n$.

Verso PCA

Dal precedente teorema è chiaro come agire:

- dato un insieme di punti nello spazio $\mathbb{R}^n$ (si veda la scheda di esercitazione)
- si calcola la loro matrice di covarianza empirica Q
- si trova il suo spettro $\mathbf{e}_1, ..., \mathbf{e}_n$, $\lambda_1, ..., \lambda_n$
- si decide che il piano individuato da $\mathbf{e}_1, \mathbf{e}_2$ è quello che fornisce la miglior visione dei punti (si veda la scheda di esercitazione: corrisponde al piano con la miglior visuale)
- L'ipotesi di gaussianità vedremo che non serve, strettamente parlando. Però ci ha fornito in modo facile e geometricamente intuitivo il risultato.

Il metodo PCA

- Data una tabella (si veda l'esempio nella scheda di esercitazione)
- le righe vengono interpretate come punti di uno spazio $\mathbb{R}^n$
- il metodo PCA calcola la matrice di covarianza empirica Q della tabella (è come se cercasse di scoprire la Q della gaussiana da cui i punti sono stati generati)
- e poi calcola la decomposizione spettrale di Q. Tutto questo col comando `princomp(B)`
- col comando `biplot` si ottiene il piano individuato da $\mathbf{e}_1, \mathbf{e}_2$ con disegnati i punti di partenza, corrispondenti alle righe ("gli individui") della tabella.

Il metodo PCA

- Quindi il metodo PCA trova le direzioni dello spazio $\mathbf{e}_1, \mathbf{e}_2$, ecc. lungo cui c'è la maggiore dispersione:
- $\mathbf{e}_1$ è la direzione con maggior dispersione (è detta *prima componente principale*)
- $\mathbf{e}_2$ è, tra le direzioni perpendicolari a $\mathbf{e}_1$, quella con maggior dispersione (*seconda componente principale*)
- e così via.
- Sono le direzioni che, via via, catturano la maggior parte della varianza. Si parla allora di *varianza spiegata* dalla prima componente principale, *varianza spiegata* dal piano principale, e così via (le definiremo tra poco).
- La varianza lungo $\mathbf{e}_1$ è λ_1, e così via.

Varianze delle componenti principali

Detto $\mathbf{X} = (X_1, ..., X_n)$, poniamo $V_1 = \langle \mathbf{X}, \mathbf{e}_1 \rangle = \mathbf{X} \cdot \mathbf{e}_1^T$, $V_2 = \langle \mathbf{X}, \mathbf{e}_2 \rangle$ e così via. (I seguenti teoremi non dipendono dall'ipotesi gaussiana, che è servita solamente per dare una maggior enfasi geometrica)

Theorem

$$\begin{aligned} Var[V_1] &= \lambda_1, ..., Var[V_n] = \lambda_n \\ Cov(V_i, V_j) &= 0 \qquad per\ i \neq j \end{aligned}$$

Definition

La varianza spiegata dalla prima componente principale (risp. dal piano principale) è

$$\frac{\lambda_1}{\lambda_1 + \cdots + \lambda_n}$$

(risp. $\frac{\lambda_1 + \lambda_2}{\lambda_1 + \cdots + \lambda_n}$).

Varianze delle componenti principali

Il teorema precedende, $Var[V_1] = \lambda_1, ..., Var[V_n] = \lambda_n$, implica che $Var[V_1] \geq \cdots \geq Var[V_n]$. Sia da questo, sia dall'immagine geometrica delle gaussiane (che però non sono più ipotizzate), risulta "chiaro" che $\mathbf{e}_1$ è la direzione di massima variabilità. Il seguente teorema lo ribadisce in modo più preciso:

Theorem

$$\max_{\|\mathbf{v}\|=1} Var[\langle \mathbf{X}, \mathbf{v} \rangle] = Var[\langle \mathbf{X}, \mathbf{e}_1 \rangle] = \lambda_1.$$

Punteggi e classifiche

Tra gli scopi di PCA c'è quello di stilare classifiche, o di attribuire punteggi, agli individui di una tabella, tramote grandezze (variabili) riassuntive che tengano conto dell'intera tabella.
Se decidiamo che la prima componente principale riassume un lato interessante della tabella, la possiamo usare a questo scopo.
Un individuo è rappresentato da un punto $\mathbf{x} \in \mathbb{R}^n$ (una riga della tabella).
Il suo punteggio rispetto alla prima componente principale è dato dalla proiezione di $\mathbf{x}$ su $\mathbf{e}_1$:

$$\langle \mathbf{x}, \mathbf{e}_1 \rangle .$$

Il comando `predict(...)` calcola tutti i punteggi rispetto a tutte e componenti principali.

Riassunto degli scopi di PCA (visti fino ad ora)

1. Vedere gli "individui" di una tabella nel modo più "aperto" possibile, così da riconoscere clusters, posizioni relative
2. Vedere a colpo d'occhio, col biplot, le associazioni tra variabili (le frecce rosse) e le posizioni relative degli individui
3. Capire se i dati sono descritti, a meno di piccolo errore, da pochi gradi di libertà
4. Trovare le direzioni di massima variabilità, ovvero i "modi" tipici con cui i dati differiscono dalla loro media
5. Misurare gli individui rispetto a variabili nuove, che siano combinazioni opportune di quelle originarie (fare classifiche rispetto a tali variabili, usare tali variabili per caratterizzare gli individui,...). Approfondiamo quest'ultimo punto.

Le variabili nascoste ed i loadings

- Nello spazio $\mathbb{R}^n$ abbiamo due basi: quella canonica, che indicheremo con $\mathbf{u}_1, ..., \mathbf{u}_n$ (definita da $\mathbf{u}_1 = (1, 0, ..., 0)$ ecc.) e quella degli autovettori di Q, ovvero $\mathbf{e}_1, ...\mathbf{e}_n$
- Corrispondentemente, abbiamo le variabili originarie $X_1, ..., X_n$
- e le nuove variabili $V_1, ..., V_n$
- Se ad es. le prime due variabili V_1, V_2 spiegano la maggior parte della varianza, allora possiamo immaginare che esse "spieghino" le $X_1, ..., X_n$, nel senso del seguente modello lineare.
- E' come se avessimo trovato delle *variabili nascoste*, le V_1, V_2, che spiegano quelle di partenza.

Le variabili nascoste ed i loadings

Dal fatto che due basi di $\mathbb{R}^n$ sono legate da opportune combinazioni lineari discende con facili calcoli che lo sono anche le variabili:

$$\begin{aligned} X_1 &= a_{11} V_1 + \ldots + a_{1n} V_n \\ &\ldots \\ X_n &= a_{n1} V_1 + \ldots + a_{nn} V_n \end{aligned}$$

dove i coefficienti a_{ij} sono detti "loadings".

Le variabili nascoste ed i loadings

Se ora isoliamo le prime due variabili V_1, V_2

$$\begin{aligned} X_1 &= a_{11} V_1 + a_{12} V_2 + \epsilon_1 \\ &\dots \\ X_n &= a_{n1} V_1 + a_{n2} V_2 + \epsilon_n \end{aligned}$$

dove $\epsilon_1 = a_{13} V_3 + \dots + a_{1n} V_n$ ecc. abbiamo trovato una relazione lineare approssimata tra le due variabili V_1, V_2 e le grandezze di partenza, che vengono quindi "spiegate" da tali variabili. Abbiamo una riduzione della dimensione del problema, del numero di variabili necessarie a descrivere il nostro fenomeno.
Gli errori sono piccoli:

$$Var\left[\epsilon_1\right] = a_{13}^2 \lambda_3 + \dots + a_{1n}^2 \lambda_n \leq \lambda_3$$

(vale $a_{11}^2 + \dots + a_{1n}^2 = 1$).

Le variabili nascoste ed i loadings

Quindi, se ad esempio avessimo un modello di regressione lineare complesso della forma

$$Y = b_1 X_1 + ... + b_n X_n + \epsilon$$

potremmo semplificarlo nella forma

$$Y = c_1 V_1 + c_2 V_2 + \epsilon'.$$

E' possibile che in qualche sempio pratico questa relazione sia più precisa del modello complesso, se le variabili nascoste in realtà catturano meglio la "fisica" del problema.
I coefficienti c_i si cacolano dai coefficienti b_i e dai loadings.
La tecnica denominata Factor Analysis è una versione più avanzata di quanto appena detto.